Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms

Dalgaty, Thomas ; Payvand, Melika ; Moro, Filippo ; Ly, Denys R B ; Pebay-Peyroula, Florian ; Casas, Jerome ; Indiveri, Giacomo ; Vianello, Elisa

Abstract: Recurrent neural networks are currently subject to intensive research efforts to solve temporal computing problems. Neuromorphic processors (NPs), composed of networked neuron and synapse circuit models, natively compute in time and offer an ultralow power solution particularly suited to emerging temporal edge-computing applications (wearable medical devices, for example). The most significant roadblock to addressing useful problems with neuromorphic hardware is the difficulty in maintaining healthy network dynamics in recurrent neural networks. In animal nervous systems, this is achieved via a multitude of adaptive homeostatic mechanisms which act over multiple time scales to counteract network instability induced via drift, component failure, or learning processes such as spike-timing dependent plasticity. One such mechanism is neuronal intrinsic plasticity (IP) where a neuron adapts its parameters which govern its excitability to fire around a target rate. The approach employed in state of the art NPs, based on a central volatile memory remotely setting model parameters, critically constrains parameter variety and bandwidth rendering realization of these essential mechanisms impossible. This paper demonstrates how reconfigurable nonvolatile resistive memories can be incorporated into neuron and synapse circuits allowing memory to be truly colocalized with the computational units in the computing fabric and facilitating the realization of massively parallel local plasticity mechanisms in neuromorphic hardware. Exploiting nonconventional programming operations of HfO2 based RRAM (stochastic SET and the RESET random variable), we propose a technologically plausible IP algorithm and demonstrate its use in the case of a recurrent neural network topology whereby the system self-organizes to sustain stable and healthy network dynamics around a target firing rate.

DOI: https://doi.org/10.1063/1.5108663

properties of RRAM for massively parallel local plasticity mechanisms. APL Materials, 7(8):081125. DOI: https://doi.org/10.1063/1.5108663

APL Materials

Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms

Cite as: APL Mater. **7**, 081125 (2019); https://doi.org/10.1063/1.5108663
Submitted: 30 April 2019 . Accepted: 29 July 2019 . Published Online: 29 August 2019

Thomas Dalgaty, Melika Payvand, Filippo Moro, Denys R. B. Ly, Florian Pebay-Peyroula, Jerome Casas, Giacomo Indiveri, and Elisa Vianello

Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms

Cite as: APL Mater. 7, 081125 (2019); doi: 10.1063/1.5108663
Submitted: 30 April 2019 • Accepted: 29 July 2019 •
Published Online: 29 August 2019

Thomas Dalgaty,[1,a)] Melika Payvand,[2] Filippo Moro,[3] Denys R. B. Ly,[1] Florian Pebay-Peyroula,[1] Jerome Casas,[4] Giacomo Indiveri,[2] and Elisa Vianello[1]

AFFILIATIONS

[1]CEA-Leti, Grenoble, France
[2]University of Zurich and ETH Zurich, Zurich, Switzerland
[3]Politecnico di Torino, Torino, Italy
[4]University of Tours, Tours, France

Note: This paper is part of the special topic on Emerging Materials in Neuromorphic Computing.
[a)]**Electronic mail:** Thomas.DALGATY@cea.fr

ABSTRACT

Recurrent neural networks are currently subject to intensive research efforts to solve temporal computing problems. Neuromorphic processors (NPs), composed of networked neuron and synapse circuit models, natively compute in time and offer an ultralow power solution particularly suited to emerging temporal edge-computing applications (wearable medical devices, for example). The most significant roadblock to addressing useful problems with neuromorphic hardware is the difficulty in maintaining healthy network dynamics in recurrent neural networks. In animal nervous systems, this is achieved via a multitude of adaptive homeostatic mechanisms which act over multiple time scales to counteract network instability induced via drift, component failure, or learning processes such as spike-timing dependent plasticity. One such mechanism is neuronal intrinsic plasticity (IP) where a neuron adapts its parameters which govern its excitability to fire around a target rate. The approach employed in state of the art NPs, based on a central volatile memory remotely setting model parameters, critically constrains parameter variety and bandwidth rendering realization of these essential mechanisms impossible. This paper demonstrates how reconfigurable nonvolatile resistive memories can be incorporated into neuron and synapse circuits allowing memory to be truly colocalized with the computational units in the computing fabric and facilitating the realization of massively parallel local plasticity mechanisms in neuromorphic hardware. Exploiting nonconventional programming operations of HfO_2 based RRAM (stochastic SET and the RESET random variable), we propose a technologically plausible IP algorithm and demonstrate its use in the case of a recurrent neural network topology whereby the system self-organizes to sustain stable and healthy network dynamics around a target firing rate.

I. INTRODUCTION

While problems in artificial intelligence regarding static data (e.g., images) have been largely solved,[1] effective processing of temporal datasets (speech, biomedical signals) remains challenging. Whereas static data are encoded in intensity, temporal data are encoded in intensity and time and therefore systems capable of extracting useful temporal features are required to retain information on the history of a data sequence. Popular approaches in feature extraction and classification of temporal data make use of recurrent artificial neural network and long-short term memory network models trained via back-propagation through time algorithms.[2] While these approaches achieve state of the art performance, their staggering training time and power consumption pose severe drawbacks for emerging edge-computing applications.[3] Spiking neural network (SNN) topologies such as recurrent SNNs

and liquid state machines (LSMs)[4] are now receiving increased attention with the promise of performing ultralow power temporal processing through emulation of the computational principles observed in animal nervous systems.[5,6] These neural network topologies, and their spike-based plasticity mechanisms, can now be emulated in an emerging class of computing system referred to as neuromorphic processors (NPs).[7,8] Neuromorphic processors interconnect analog or digital neuron and synapse circuit models, intended to emulate neural dynamics, in a reconfigurable manner allowing neural networks to be realized in a highly parallel computing system. NPs utilizing analog circuit models boast the lowest power consumption and consequently are the most suited for emerging ultralow power edge-computing applications. State of the art analog NPs typically use centralized volatile memories to set parameters of the distributed neuron and synapse models. However, this approach poses severe drawbacks and constrains state of the art NPs as will be described in Sec. V. Furthermore, it has been demonstrated that in order to maintain healthy dynamics in recurrent neural networks, therein dynamics that permit effective computation, a variety of adaptive homeostatic plasticity mechanisms are required.[9,10] These homeostatic mechanisms counteract sources of network instability arising from drift, component failure, or learning processes such as spike-timing dependent plasticity which could result in networks becoming excessively excited or inactive. One such mechanism is neuronal intrinsic plasticity (IP) whereby a neuron adapts its excitability to fire around a target firing rate.[11,12] In analog NPs, realization of such mechanisms over large time scales and network sizes is extremely challenging, resulting from severe constraints imposed by the technology. In this paper, we propose that hybrid neuromorphic circuit models, which incorporate nonvolatile resistive memories (RRAM) into CMOS circuits, can solve substantial problems facing NPs in lack of parameter variety, power consumption, temperature instability, and the implementation of the massively parallel local neural and synaptic plasticity mechanisms. Specifically, we demonstrate how to incorporate HfO_2 based one transistor one resistor (1T1R) RRAM structures into a differential pair integrator (DPI) neuron circuit and a DPI synapse circuit. We then show how measured, nonconventional properties of the memory's RRAM SET (stochastic SET) and RESET (random variable RESET) programming operations can be exploited by further local circuits to realize massively parallel local plasticity mechanisms—such as neuronal intrinsic plasticity. Finally, we show in a spiking neural network simulation that a recurrent spiking neural network topology, composed of hybrid DPI neurons (employing the proposed algorithm) and DPI synapses, can self-organize and fire ensemble around a target rate.

II. HYBRID NEUROMORPHIC CIRCUITS

The basis of hybrid neuromorphic circuits is the 1T1R structure[13,14] depicted in Fig. 1(a). A resistive memory (R1 or R2) is connected in series with either a PMOS or a NMOS selector transistor (T1 or T2). The transistor has two roles: (1) to determine the share of total programming voltage $V_{top} - V_{bot}$ (V_{prog}) that is seen over the resistive memory and (2) to limit the current flowing through the device during a programming operation. Both objectives are achieved by modulating V_{gate} when a nonzero V_{prog} exists.

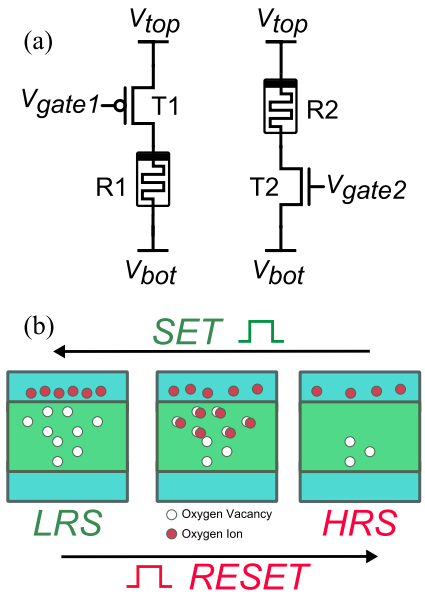

FIG. 1. The one-transistor-one-resistor structure and the mechanism of resistance modulation. (a) 1T1R circuit schematic. (b) Oxygen vacancy based working principle of the two restive states.

There are two standard RRAM programming operations called SET and RESET and two resulting memory states called the low (LRS) and high (HRS) resistive states. For the case of oxide-based RRAM (OxRAM) [Fig. 1(b)], a thin layer (tens of nanometers) of a transition metal oxide (TMO) material is sandwiched between two metal electrodes and can have its resistance modified through application of electrical pulses. The resistance of the TMO depends largely on the number of oxygen vacancies which are created or removed through voltage induced reduction-oxidation (REDOX) reactions with the electrodes. In the case of bipolar OxRAM, a positive V_{prog} (V_{set}), applied to the top electrode, creates an oxygen-poor conductive filament through which electrons can flow. This positive voltage pulse is a SET programming operation which puts the device into the LRS. This oxygen-vacancy based conductive filament can thereafter be disrupted with application of a negative V_{prog} (V_{reset}) voltage pulse in a RESET operation flipping a device into the high resistive state. This is normally achieved through application of a positive pulse to the bottom electrode. In traditional memory applications, the LRS and HRS are used to represent a binary 1 or 0 by means of resistance thresholding. Unlike volatile memory technologies, the memory state persists in the absence of a power supply and is therefore referred to as nonvolatile memory (NVM).

Ion channels within neuronal membranes regulate the flow of ionic current into and out of the cell's somatic body which acts as a capacitor. Essentially, they represent transient or fixed resistances which regulate flow of charge between an extracellular battery and this capacitor and serve as a fundamental building block of animal nervous systems. In the same fashion, we propose that (volatile[15] and nonvolatile[14,16]) resistive memory technologies, as a parallel to ion channels, can serve as the fundamental building blocks

in the construction of artificial hybrid neuromorphic computing systems.

A. Hybrid DPI neuron

The most straightforward, yet still computationally useful, neuron models are the leaky-integrate and fire (LIF) models. They capture the essence of a neuron's ability to integrate charge on its somatic membrane upon synaptic excitation while simultaneously leaking away this charge in time. Furthermore, upon reaching a threshold of accumulated charge, the neuron fires and emits an output pulse which can be propagated to the synaptic inputs of other LIF model neurons. The hybrid differential pair integrator neuron model in Fig. 2(a) captures these behavioral features in an hybrid CMOS-RRAM circuit. Upon the injection of input current, charge is integrated onto capacitor C1. The amount of integrated charge depends on the ratio of the resistance values $1T1R_2$ (green) to $1T1R_1$ (blue) and therefore allows for gain tuning. The charge which is integrated onto C1 leaks to ground at a rate defined by the resistance of $1T1R_2$ (green). If the rate of integration sufficiently exceeds the rate of the leak, then a threshold voltage is reached (V_{th1}) (here defined using an OPAMP comparator) and an output inverter sets V_{out} to a logic high. During this firing event, capacitor C2 is charged via the now open current source M4. As soon as the capacitor exceeds V_{th2}, transistor M5 opens and shunts V_{in} to ground—bringing to an end the pulse. Transistor M5 remains shunted to ground for the period the voltage on capacitor C2 remains in excess of V_{th2}, defined by the rate the charge leaks to ground through $1T1R_3$. The RRAM $1T1R_1$ and $1T1R_2$ affect the neuron input time constant and input gain, while $1T1R_3$ defines the neuronal refractory period. The effect of each of the individual resistances was studied in Ref. 17. Two waveforms with different resistance configurations, obtained through SPICE simulation, plot V_{in} and V_{out} under a periodic current spike train (1 μs pulse-width of 100 nA every 250 μs), are shown in Fig. 2(b).

B. Hybrid DPI synapse

While the input currents in Fig. 2(a) were simple pulses, the synaptic currents injected into neurons in biology exhibit temporal properties which are important for neural computation.[18] Circuit models exist for mimicking synaptic dynamics for use in neuromorphic processors.[19] The simplest model is that of the exponential synapse whereby, during an input voltage pulse (modeling a presynaptic action potential), the output current is stepped and then decays exponentially in time. This is the behavior of the hybrid differential pair synapse circuit in Fig. 3(a). Upon a V_{in} pulse, a current proportional to the value of $1T1R_2$ (green) flows from C1 to ground. As this current flows, during an active high V_{in} pulse, the voltage at C1 reduces and turns on transistor M3, allowing an output current to flow (which can be injected into a neuron circuit model). This voltage over C1 continues to reduce for as long as the voltage difference between C1 and the potential divider node between $1T1R_1$ and $1T1R_2$ is large enough to keep the diode connected transistor M1 turned on—therefore, $1T1R_1$ imposes a limit on the magnitude of the output current. After an input pulse comes to an end, so does the reduction of the voltage over C1, and instead, the capacitor charges up again linearly via a leakage current from $1T1R_3$ (red). This results in an exponential reduction in the output current. A SPICE simulation in Fig. 3(b) gives two examples of the output current waveform after an input voltage pulse for two configurations of the three 1T1R

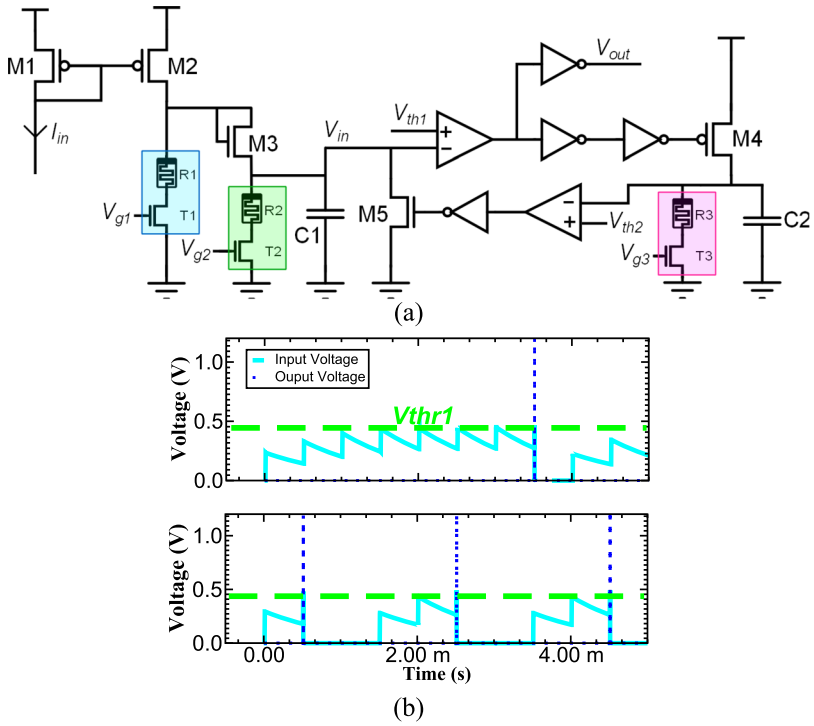

FIG. 2. The hybrid differential pair integrator neuron circuit and its behavior. (a) Hybrid differential pair integrator neuron circuit where 1T1R structures are used to set the input gain, time constant, and refractory period. All NMOS transistors have a width/length of 650 nm/250 nm, while the PMOS transistors are 1.2 μm/250 nm. Capacitors C1 and C2 are both 1 pF. (b) The circuit is stimulated with a train of square current pulses (1 μs pulse-width of 100 nA every 250 μs) where the input voltage, output voltage, and input current are plotted. The supply voltage is 1.2 V consistent with the voltage rating for the 130 nm CMOS technology used in simulation. Two resistance configurations are presented which are (top) R1 = 1 GΩ, R2 = 1 GΩ, R3 = 1 GΩ, and (bottom) R1 = 40 MΩ, R2 = 1 GΩ, R3 = 40 MΩ.

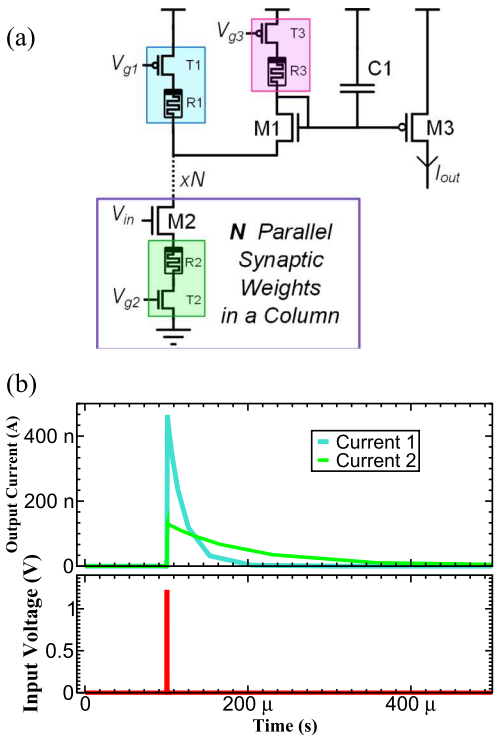

FIG. 3. The hybrid differential pair integrator synapse circuit and its behavior. (a) Hybrid differential pair integrator synapse circuit where 1T1R structures are used to set a weight per synapse and current gain and the time constant of the exponential current decay per column in a synaptic array. All NMOS transistors have a width/length of 650 nm/250 nm, while the PMOS transistors are 1.2 μm/250 nm. Capacitor C1 is 1 pF. (b) During an input voltage pulse (red pulse of 1.2 V with a 1 μs pulse-width), the output current is incremented. After the pulse, the current exponentially decays to zero. The current waveform for two different configurations are shown—current 1 (R1 = 10 MΩ, R2 = 500 kΩ, R3 = 1 GΩ) and current 2 (R1 = 10 MΩ, R2 = 375 kΩ, R3 = 500 MGΩ). The supply voltage is 1.2 V consistent with the voltage rating for the 130 nm CMOS technology used in simulation.

structures which augment the hybrid circuit. It should be noted that, although in Fig. 3(a) one synapse circuit contains one capacitor, inside neuromorphic processors[7] multiple synapse circuits share (along a row or column of a synaptic array) a capacitor and superimpose their currents onto it. This helps reconcile the small footprint of a synapse circuit with the large footprint of a 1pF capacitor without compromising on large (biological) time constants.

III. NONCONVENTIONAL PROPERTIES OF HfO_2 BASED RRAM

HfO_2 based RRAM are conventionally used as binary devices switching between a low and a high resistance state in a deterministic way for standard memory applications. Here by contrast, we would like to treat the SET operation as a stochastic process using a subthreshold V_{prog}. In addition, we view, as a result of the HRS cycle-to-cycle variability, the RESET operation as a random variable conditioned on V_{prog} and V_{gate}. These real device properties can be used to develop technologically plausible neuromorphic and in-memory computing stochastic algorithms, such as in-memory Markov processes.[20] The stochastic SET and RESET random variables of HfO_2 based RRAM 1T1R structures with Ti/TiN electrodes, integrated monolithically in 130 nm CMOS process,[14,16] are characterized in this section. A scanning electron microscope image of a wafer cross section, with CMOS and HfO_2 based RRAM on the same substrate, is shown in Fig. 4 where the memories have been deposited between metal layers 4 and 5 in the back-end-of-line and can be interfaced to CMOS circuits in the front-end-of-line through vias between metal layers 3, 2, and 1.

A. Stochastic SET

Traditionally, a SET programming pulse is applied which ensures with certainty that a functioning device transitions from the HRS to the LRS. However, for the case of subthreshold SET pulses (here below V_{set} = 1.4 V), the HfO_2 based RRAM exhibits a nondeterministic switching mechanism whereby the probability of a device being SET has a dependence on the SET voltage applied over the device.[21] In order to characterize this probability-voltage relationship, devices in a 4 kbit (16 × 256) 1T1R matrix were subject to a sweep of subthreshold SET pulses (devices were reinitialized to an initial HRS state between V_{set} steps). A resistance threshold of 20 kΩ defines a SET device from the one which remains in the HRS. The fraction of SET devices after the subthreshold SET pulse had been applied defines the SET probability per voltage across the matrix. The cumulative distributions (CDFs) of the 4096 devices in the matrix for a sweep of V_{set} are plotted in Fig. 5(a). As V_{set} increases, devices are more likely to transition from the HRS distribution (right) to the LRS distribution (left). Furthermore, it is interesting to note that even for deep subthreshold pulses the resulting LRS resistance values fall under the 20 kΩ threshold and into the LRS distribution despite a small (relative to standard SET conditions) applied programming voltage. The probability extracted at each V_{set} is plotted in Fig. 5(b) for 3 different pulse-widths (100 ns, 500 ns, and 10 μs). The probability-voltage relationship is seen to be

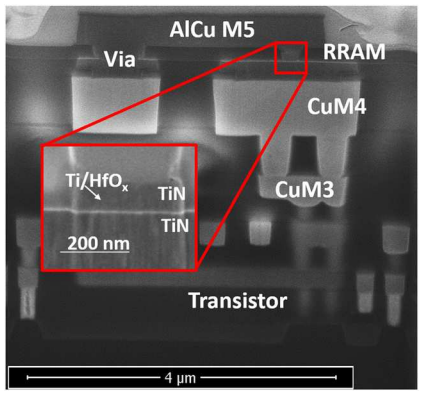

FIG. 4. HfO_2 based RRAM is integrated monotonically in the back-end-of-line with CMOS transistors in the front-end-of-line in a 130 nm CMOS process. This process permits design of hybrid circuits where RRAM can coexist physically above CMOS circuits on the same chip. The HfO_2 thin film is shown zoomed within the inner red box, and the location of the CMOS is marked by the text transistor.

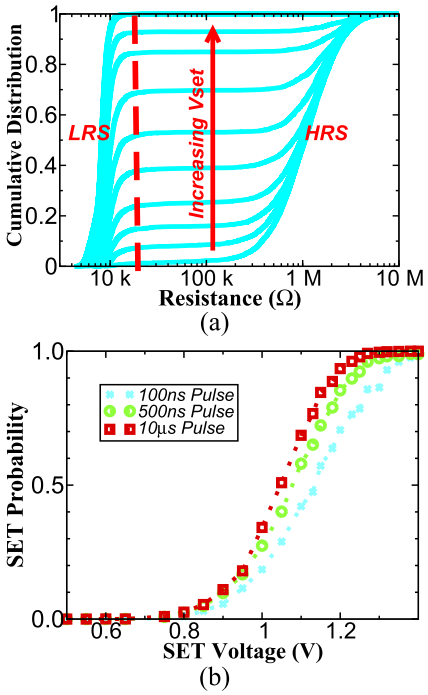

FIG. 5. Viewing the SET operation as a stochastic process, there is a sigmoidal relationship between SET probability and SET voltage for subthreshold (less than 1.4 V) voltage pulses. (a) Cumulative distributions of device states after subthreshold programming pulses for a sweep of SET voltages. Those assuming a resistance below a 20 kΩ threshold are defined as SET in the low resistive state. (b) The probability of setting a device has a sigmoidal dependence on the voltage of the SET programming pulse where the slope of the sigmoid can be increased using longer SET programming pulses.

sigmoidal where a small degree of control in the slope of the sigmoid can be exerted by varying the pulse-width.

1. Intercycle/cell variability

The sigmoidal relationship between the SET voltage and the corresponding switching probability (verified across multiple dies and wafers) describes well the properties of the stochastic SET for a population of memories. In the case of the hybrid circuits, single structures are integrated into single cells, and therefore, it becomes important to understand the variability in the switching probability between single devices. In order to characterize this, 100 cycles of subthreshold SET operations were performed with a subset of V_{set} voltages. The switching probability, per device, corresponds to the number of times it was SET over the 100 cycles. The deviation between the probability of a single device and the mean probability (the mean of all devices over 100 cycles) is plotted for three mean probabilities in a heatmap in Fig. 6. Soft reds and blues correspond to devices with switching probabilities equal to or close to the mean per applied SET voltage. Stronger reds and blues indicate, by contrast, devices which have a switching probability significantly less (blue) or greater (red) than this mean. It is clear from visual inspection that a substantial device-to-device (D2D) variability in the switching probability is present. This D2D variability is captured more explicitly using a boxplot in Fig. 7. Here, the median (blue horizontal line), ±25% percentile (red box), ±50% percentile (red whiskers), and ±95% percentile (blue points) are plotted. The dispersion is most pronounced at voltages corresponding to probabilities between 0.2 and 0.8. For example, for V_{set} = 1 V, the median probability is approximately 0.6, but half of the device population, defined by the limits of the purple box, exhibit SET probabilities between 0.3 and 0.85. The NIST test suite SP800-22[22] was used in order to evalaute if a spatial correlation in the D2D SET probability existed across the matrix. This test suite is commonly used to validate random number generators by running 15 tests on the generator output, especially searching for spatial correlations. The number walk, composed of the complete 4 kbits of the matrix over 100 cycles, passes the full suite of tests. According to these tests, the D2D spatial correlation can be confidently considered as nonsignificant.

B. RESET random variable

The objective in a standard RESET operation is to switch the device to the HRS (from the LRS) such that the resulting resistance state is comfortably above the resistance threshold while also maximizing the device endurance. Unlike the abrupt nature of the

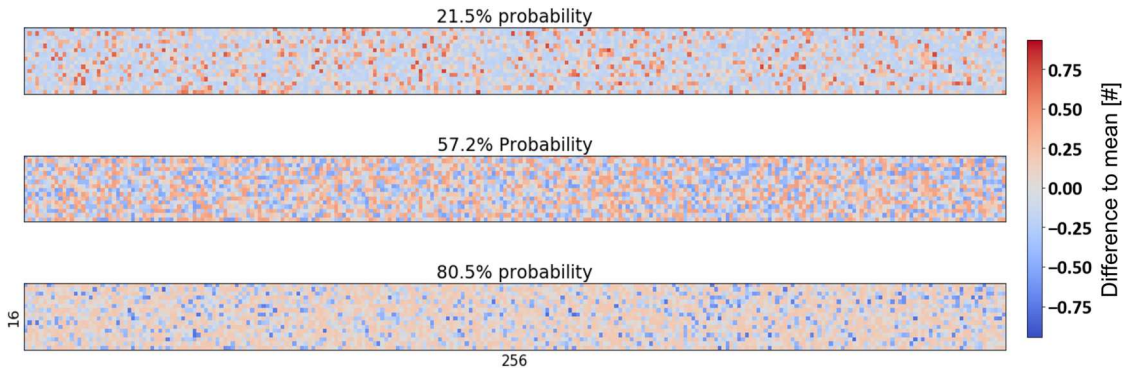

FIG. 6. Switching probability for each device in a 16 × 256 cell 4 kbit matrix colored based on its deviation from the mean switching probability over a 4 kbit 1T1R matrix. Strong reds and blues indicate a significant deviation, while softer shades show cells close to the mean probability.

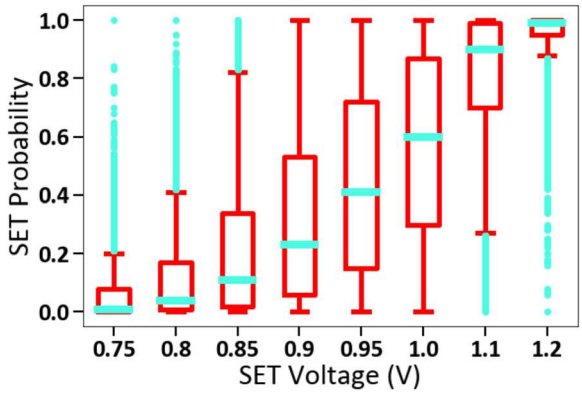

FIG. 7. Device to device variability captured with a boxplot. Here, the median (blue horizontal line), ±25% percentile (red box), ±50% percentile (red whiskers), and ±95% percentile (blue points) are plotted.

SET operation, the RESET is a gradual process[23] where the resistance becomes greater with consecutive RESET pulses. Also, unlike in the SET, the HRS resistance is strongly influenced by the value of V_{reset} and V_{gate}. Therefore, although often done, it is artificial to extract a RESET probability-voltage relationship. However, this does not say that by any means the RESET operation is deterministic. On the contrary, the process governing the oxygen-vacancy filament dissolution is clearly also random as observed in the cycle-to-cycle (C2C) variability in the HRS resistance value (for identical programming conditions). Therefore, due to this inherent C2C variability, the RESET operation in HfO_2 based RRAM can be viewed as sampling from a probability distribution (PDF) and therein treated as a random variable. The relationship between V_{reset}, V_{gate} and the C2C HRS distribution (mean resistance and two standard deviations), obtained with 100 cycles on a single device, is plotted in Fig. 8(a). The gate voltage has the effect of limiting the HRS PDF mean resistance for an increasing RESET voltage. Before this saturation, there is a clear region where mean C2C resistance can be controlled with the applied programming voltages. In the case of V_{gate} = 4 V, HRS resistances span a range of 5 orders of magnitude with the highest values, using the strongest measured conditions (V_{reset} = 4 V and V_{gate} = 4 V), slightly below 1 GΩ. In the context of hybrid neuromorphic circuits Fig. 2(a), this translates as being able to vary neural time constants over 5 orders of magnitude and, assuming capacitors on the order of pF, permits neural time constants in the millisecond regime to be obtained. For applications addressing real-time problems in a natural environment, it is essential that the time constant of network dynamics and the environment be matched whereby many environmental processes have time constants on the order of milliseconds. It should be noted that for strong RESET programming conditions, the endurance of the devices degrades significantly. In order to better define the distribution shape, a single device was cycled 1000 times at two RESET conditions [V_{gate} = 4 V and V_{reset} = 2 V (red) and V_{gate} = 4 V and V_{reset} = 1.5 V (blue)] in Fig. 8(b). Consistent with previous results,[16] the HRS C2C probability density can be well described by a log-normal distribution, as in Fig. 8(b). Therefore, the RESET operation can be viewed, specifically, as a

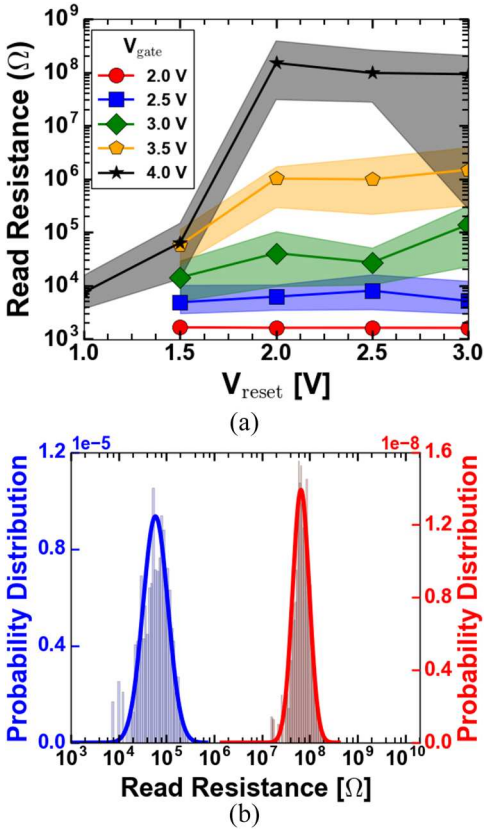

FIG. 8. Treating the RESET high resistive state cycle to cycle variability as sampling from a log-normal random variable and the dependence on RESET programming conditions. (a) The mean (data points) and the spread at two standard deviations (shaded region) of the HRS resistance state are plotted for a range of V_{reset} for a sweep of V_{gate}. (b) Two examples of HRS distributions for different RESET programming parameters [V_{gate} = 4 V and V_{reset} = 2 V (red) and V_{gate} = 4 V and V_{reset} = 1.5 V (blue)]. The distributions can be well fitted using a log-normal probability distribution where the standard deviation of the underlying normal distribution is between 0.4 and 0.5.

random variable where the PDF is a log-normal distribution with a mean conditioned on V_{gate} and V_{reset} during a RESET operation. The standard deviations (of the underlying normal distribution to the log-normal) were extracted and found to be between 0.4 and 0.5, inline with measured dispersion for the same technology.[24] Note that previous results have demonstrated an additional influence of the recent history of HRS states on the current state for weak programming conditions (low V_{reset}), whereby a correlation exists over the course of tens of cycles.[25]

IV. TECHNOLOGICALLY PLAUSIBLE INTRINSIC PLASTICITY

Intrinsic plasticity has proven essential in maintaining healthy dynamics in recurrent neural networks.[9] However, to map and export such algorithms onto state of the art neuromorphic processors is not currently technologically plausible resulting

from the constraints detailed in Sec. V. Technological plausibility demands distributing nonvolatile memory throughout the computing fabric such that, like in biology, memory and computation are colocalized and indistinguishable. We have shown how this can be achieved using hybrid neuron and synapse circuit models. We also characterized nonconventional computational properties of oxide-based RRAM that can be exploited in implementing stochastic algorithms. In this section, we outline a technologically plausible intrinsic plasticity algorithm, based on these properties, and evaluate its performance.

A. Algorithm

Intrinsic plasticity requires that individual neurons self-organize to fire around a target rate.[9] We propose that a neuron can measure its own firing rate and, at fixed intervals (here 400 ms), compare this rate with a target and, based on this difference, perform SET/RESET cycles on $1T1R_1$ and $1T1R_2$ of the hybrid DPI neuron [Fig. 2(a)]. These parameters control the input gain and input time constant and thus determine the neuron excitability. Since after every RESET operation the RRAM resamples its resistance value, the behavioral properties of the neuron will change accordingly. The algorithm is depicted in Fig. 9. We propose to periodically generate SET voltage pulses with an amplitude as a function of the firing rate difference, directly over the neuron's incorporated 1T1R structures. This exploits their inherent switching probability-voltage dependence [Fig. 5(b)] to make a decision on whether to resample their resistance values or not. Circuits have been previously described that allow SET voltage pulses to be a precisely controlled function of the firing rate difference.[17,26] This allows for the resampling probability sigmoid to be a function of the firing rate difference and also for the sigmoid function properties (horizontal shift and slope) to be artificially augmented to realize a probability-error (error between the target and measured rates) sigmoid. A tolerance can be introduced for example. This tolerance sets a minimum error between the target rate and the measured one that is tolerated before the resampling probability for a neuron becomes nonzero. The tolerance is an important quantity in the algorithm. A value too small will prevent convergence to a stable state, since the neuron parameters will be highly sensitive to small fluctuations in activity. At the other extreme, an excessively large tolerance would prevent a neuron from organizing itself at all. Additionally, the relationship for overfiring and underfiring can also be set independently. We propose that $1T1R_1$, since it impacts only the gain, should resample from an HRS PDF with a mean equal to its current resistance value, while $1T1R_2$ should resample from an HRS PDF with a mean shifted by a constant learning rate from its previous value. The learning rate multiplied by the current resistance value and then added to or subtracted from this value gives the value of the new mean. Since $1T1R_2$ has a positive correlation with the firing rate (as it governs the input time constant), this mean shift should be positive for underfiring and negative for overfiring.

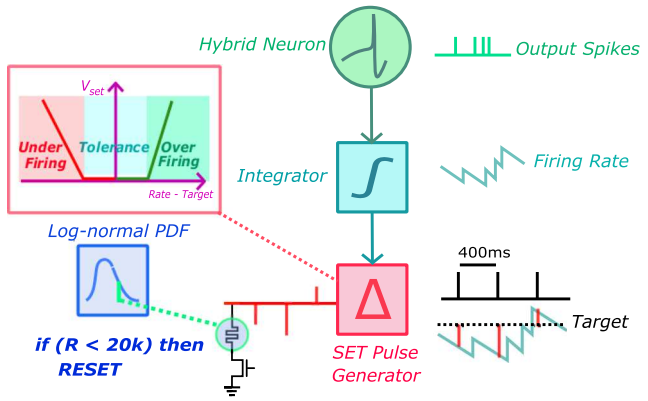

FIG. 9. Diagram of the proposed intrinsic plasticity algorithm. The hybrid DPI neuron has two RRAM 1T1R that set the properties of the neuron model (green circle). The neuron propagates a spike/pulse train to an integrator circuit (light blue block) which transforms the discrete voltage pulses into a continuous analog voltage encoding its activity. This signal (blue waveform) is compared with a target (black dashed line), and periodically (black pulse train), the error is evaluated (red pulses). Based on these differences, SET voltage pulses are generated (red block) over the incorporated 1T1R structures in their high resistive states. This intrinsically makes a stochastic decision on whether its resistance value should be resampled. If the device is SET, the resistance is below 20 kΩ, and then it is immediately RESET at which point the resistance values of the neuron memories are resampled from a log-normal PDF (navy blue block) corresponding to the inherent probability density of the HRS C2C resulting from a RESET operation. Previous work has shown that the pulse generator can control the applied SET voltage for a given error[17] between the target and measured rates. This also allows a tolerance to be introduced whereby a specified level of error is tolerated before the resampling probability becomes nonzero.

B. Recurrent neural network with hybrid IP neurons

Spiking recurrent neural network topologies mapped onto neuromorphic processors will be essential in effectively solving emerging low-power temporal edge-computing problems. Current neuromorphic processors will struggle to meet the requirements of such applications since they cannot implement the local, massively parallel plasticity mechanisms, such as neuronal intrinsic plasticity, required to obtain and sustain healthy recurrent network dynamics. In this section, we demonstrate, through spiking neural network simulation, the effect of the proposed algorithm on the topology illustrated in Fig. 10. In this topology, an input layer (blue) of 12

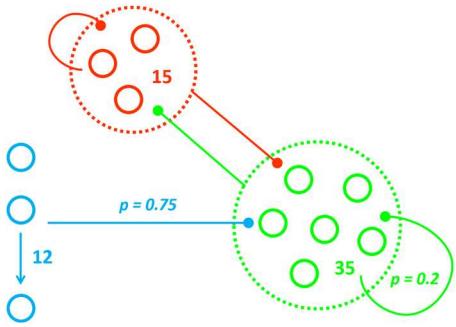

FIG. 10. The recurrent spiking neural network topology used in simulation. An input Poisson group (blue) feeds forward to an excitatory neuron population which are hybrid DPI neurons with intrinsic plasticity (green). A population of inhibitory neurons (red) is excited by the excitatory population and feeds back with inhibiting synapses that impose an upper limit for the mean firing rate in the excitatory population.

Poisson neurons[27] feed-forward into a recurrently connected excitatory population of 35 neurons (green) with a connection probability of 0.75. Poisson neurons are neurons which fire at random intervals such their interspike time PDF is a decaying exponential function. The excitatory neurons have a 0.2 chance to connect recurrently amongst themselves. There is no spatial connectivity kernel as is the case for LSMs.[4] In addition, the excitatory population excites an inhibitory population (red). The neurons in this population recurrently connect amongst themselves and also project inhibitory synapses to the excitatory population—putting on the brakes via negative feedback when it is excessively excited. The neurons in the excitatory population are equipped with intrinsic plasticity. The tolerance is set to 70 Hz for both overfiring and underfiring, while the learning rates for $1T1R_2$ were 0.05 and 0.3 for overfiring and underfiring, respectively. All synapses are the hybrid DPI synapses of Fig. 3(a), the neurons in the excitatory population are hybrid DPI neuron models [Fig. 2(a)], while the inhibitory population are simply LIF neuron models. The resistance values of the hybrid neurons are bounded within the order of the measured values in Fig. 8(a).

First, for illustrative purposes, the mean firing rate and standard deviation in the firing rate for the 35 excitatory neurons are plotted in the absence of an IP algorithm in Fig. 11(a). The mean rate oscillates around a natural frequency of 200 Hz, while the standard deviation amongst firing rates within the population is 50 Hz. By contrast, Fig. 11(b) plots the same metrics for a single run of the simulation where the neurons in the excitatory population employ the proposed IP algorithm—given a target of 120 Hz. After an initial transient period of excessive firing, the network self-organizes in 5.5 s and then settles in a configuration where the mean firing rate respects the stipulated target. The standard deviation amongst the firing rates is 38 Hz. Finally, in Fig. 11(c), the number of SET/RESET programming cycles (during each 400 ms refresh) drops from an initial count of 34 cycles to 2.1 cycles. Low RRAM switching activity is an equally important indication of convergence since not only should the network mean tend to the target (while maintaining an acceptable standard deviation amongst the individual rates in the population) but the switching activity should also cease (or become negligible). The HRS C2C log-normal standard deviation (of its underlying normal distribution) was set to 0.5 [as measured in Fig. 8(b)], while it was assumed that the D2D variability in the stochastic SET was zero (which was of course measured not to be the case). The effect of the D2D SET variability is evaluated in Sec. IV B 1. The performance of the network can be described by the three performance metrics which are annotated in Figs. 11(b) and 11(c)—time to convergence (T), standard deviation amongst firing rates after convergence (B), and number of SET/RESET cycles after convergence (C).

1. Impact of device variability

The two nonconventional RRAM programming operations come with inherent variability. The C2C variability in the HRS after a RESET operation corresponds to the standard deviation (of the underlying normal distribution) of a log-normal PDF, while the D2D variability in the stochastic SET has the effect of an undesired horizontal shift of the probability-error sigmoid (therefore impacting the tolerance and horizontally shifting it from the intended

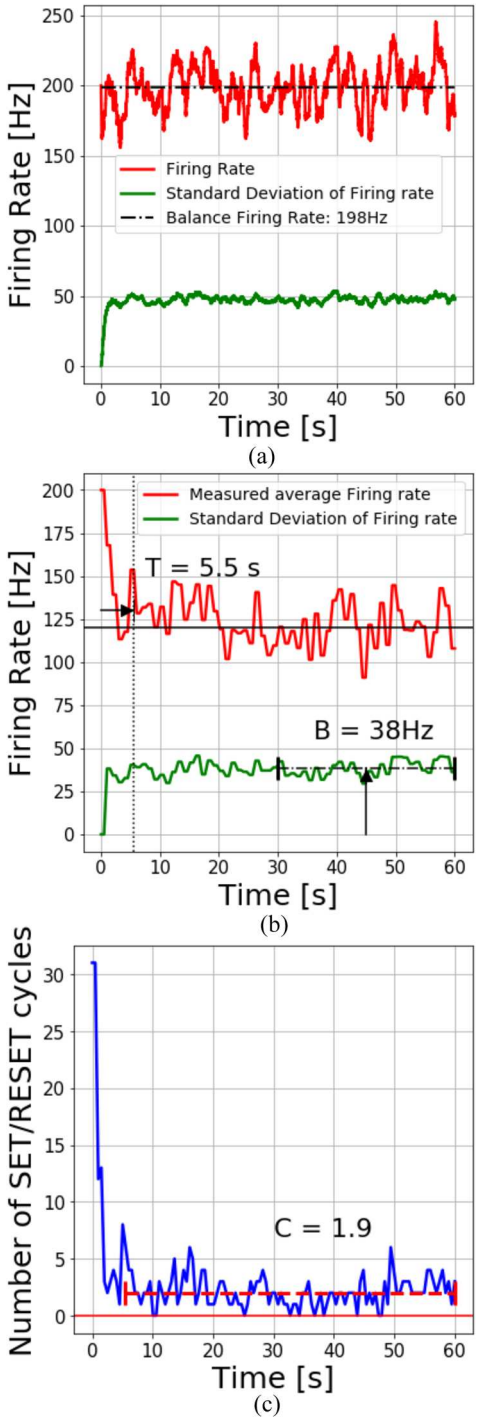

FIG. 11. Spiking neural network simulation. Intrinsic plasticity allows the recurrent network to self-organize to fire at a target rate while also minimizing standard deviation and SET/RESET resampling cycles. (a) RNN of neurons without IP firing at their natural frequency. (b) RNN of IP neurons co-organizing to fire around the target in 5.5 s in this single simulation. The standard deviation settles to 38 Hz after convergence. (c) After convergence, the number of SET/RESET cycles drops to near zero.

value). We determine their effect by using the defined performance metrics [annotated in Figs. 11(b) and 11(c)—time to convergence (T), standard deviation amongst firing rates after convergence (B), and count of SET/RESET cycles after convergence (C)] averaged over 3 independent runs from randomly initialized parameters. We first examine the impact of the C2C HRS variability on the network in Figs. 12(a) and 12(b). It is seen for low values of standard deviation of the C2C HRS PDF that the network struggles to settle to a mean precisely equal to the target—although it has a low standard deviation amongst firing rates and low count of SET/RESET cycles after converging. This is likely linked to the result of Fig. 12(b) that the convergence time drops for a higher C2C HRS standard deviation up to 0.5 and suggests that due to the wider lognormal PDF the network is able to explore a wider range of resistance values faster. However, for values larger than 0.5, the convergence time is then seen to increase. Most likely, this results from the C2C HRS PDF becoming too wide, and the step taken from the previous configuration is too large and not sufficiently correlated. With a lower C2C HRS standard deviation, the network resistances change more gradually; hence, when the network arrives to a mean rate within the specified tolerance, the memories stop resampling their parameters before reaching the exact target. The measured value of the C2C HRS standard deviation (of the underlying normal distribution) was between 0.4 and 0.5. Within this range, the algorithm is seen to find a sweet spot and converge significantly faster than values higher or lower. This leads to the conclusion that the intrinsic C2C HRS variability has a positive impact on performance. In Figs. 12(c) and 12(d), the same metrics are plotted but for the case of D2D SET variability. Here, we sample undesired horizontal shifts in the probability-error sigmoid from a normal distribution for each device. Therefore, each has a permanent offset from the desired value which impacts the effective tolerance. The time to convergence in Fig. 12(d) increases with a greater standard deviation of the normally distributed shifts in the probability-error sigmoid. However, the standard deviation amongst the neuron firing rates, the mean distance from the target firing rate, and the count of SET/RESET cycles appear to be largely unaffected. This result is encouraging since it appears that, even in the presence of significant D2D variability, the IP algorithm allows the network to self-organize and find a configuration which can compensate for the nonideal devices and fire around the target at the expense of a longer period of self-organization.

2. Power consumption

A SET/RESET cycle, required to resample a parameter, incurs a fixed penalty in energy, and therefore, such an algorithm will consume an amount of energy proportional to the update rate (here 400 ms) and the number of devices in a network which have undergone a SET/RESET cycle during this periodic update. Under standard programming operations (SET: V_{set} = 2 V, V_{gate} = 1.3 V and RESET: V_{reset} = 3 V, V_{gate} = 3 V, both with a programming pulse-width of 100 ns), the 1T1R structures studied in this paper consume approximately 50 pJ per SET/RESET cycle. Neurons also pay an energy penalty every time they spike (for the DPI neuron in 180 nm CMOS, this is 800 pJ). This is therefore an order of magnitude more expensive than a SET/RESET cycle and approximately two orders of magnitude more frequent. Clearly, as is the case in biology,

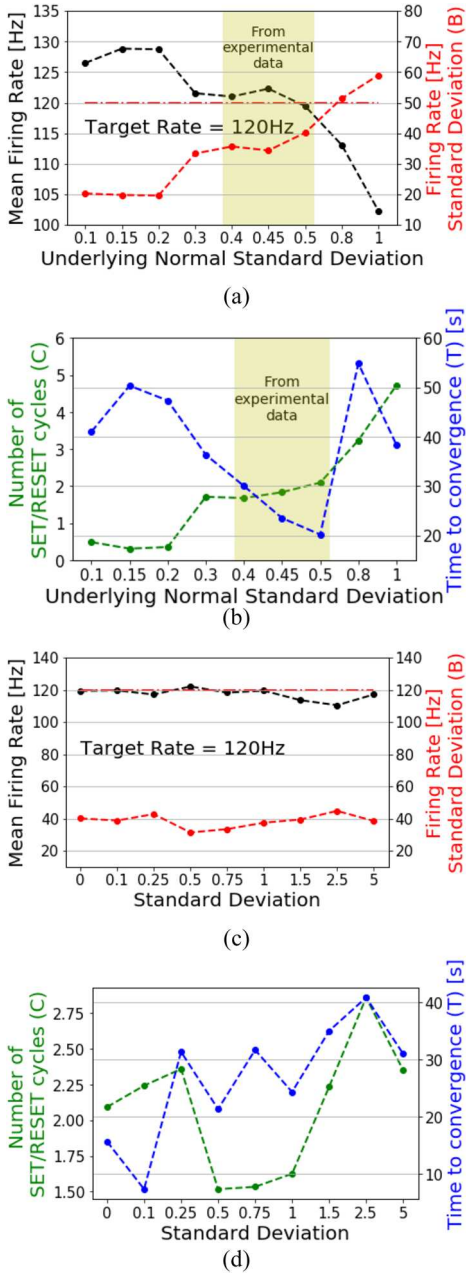

FIG. 12. The impact of cycle-to-cycle variability in the RESET and device-to-device variability in the subthreshold SET on performance metrics of the intrinsic plasticity algorithm acting on the recurrent spiking neural network are studied. (a) Impact of the standard deviation (of the underlying normal) in the cycle-to-cycle high resistive state resistance log-normal probability density function (following a RESET) on the mean firing rate and standard deviation in the firing. (b) Impact of the standard deviation (of the underlying normal) in the cycle-to-cycle high resistive state resistance log-normal probability density function (following a RESET) on convergence time and the number of SET/RESET cycles after convergence. (c) Impact of normally distributed device-to-device SET probability standard deviation on firing rate and standard deviation in firing rate. (d) Impact of normally distributed device-to-device SET probability standard deviation on convergence time and the number of SET/RESET cycles.

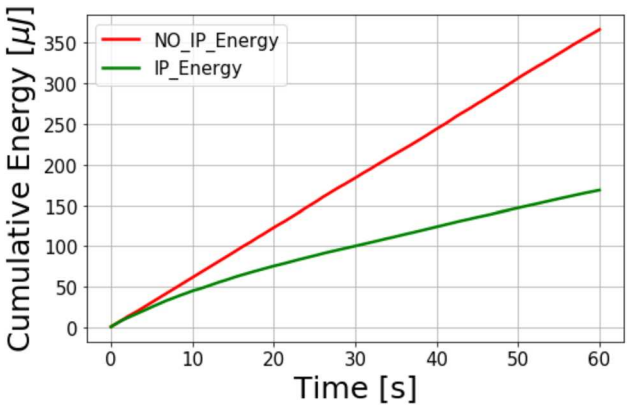

FIG. 13. The cumulative energy consumed by a recurrent neural network firing at its natural frequency (red) and the same network implementing the described intrinsic plasticity algorithm (green).

it becomes advantageous to expend a small amount of energy to reduce the (comparatively) much greater energy consumed via neural activity. As an illustrative example, we plot the cumulative energy consumption of the two networks in Fig. 11(b) (one employing IP and the other firing at its natural rate) in Fig. 13. Since the target firing rate (120 Hz) is significantly lower than the natural rate (200 Hz) the energy consumed, despite the cost of initial organization, is reduced by half. This demonstrates the opportunities in energy management of the algorithm in applications in which the system is not connected to a reliable source of power—wearable medical devices in between charging, for example.

V. ADVANTAGES OF HYBRID SYSTEMS

The dynamics of a neural network are set by the parameters of its neurons and synapses. These parameters can include, for example, the integration time constant for the synaptic dynamics and neurons, neural refractory period, synaptic efficacy, and neuron's gain and adaptation time constant.[19] In state of the art mixed-signal neuromorphic processors, such parameters are stored digitally in registers inside bias generator blocks which control the bits of current digital to analog converters (DACs) which in turn propagate voltages to bias transistors inside the neuron and synapse circuit models. By contrast, what we have proposed in this paper decentralizes the memory from the volatile digital programmable bias generators by distributing nonvolatile memories throughout the computing fabric such that they are incorporated into the neuron and synapse circuit models themselves. The benefits of our approach are multifold:

- The bias generator block burns static power which grows with the number of parameters allowed to be on the chip. In hybrid systems, the static power consumption reduces to zero as the transistor biases are replaced by incorporated, passive resistive memories.
- State of the art NPs are often forced to compromise on parameter variety such that all of the neurons on a core are obliged to share the same model parameters. If parameters were not shared, the static power consumption and the area consumed by wires (metal lines) running across the chip for connecting the biases explode with the number of parameters. For hybrid systems, each circuit model has its own parameter set by the incorporated RRAM without area or static power overhead. Such an approach also enables the self-organization of individual parameters locally.
- The effect of transistor mismatch is highly detrimental in subthreshold CMOS circuits and therefore in NPs. Since each neuron and synapse model can be individually configured, the models can locally compensate for the mismatch present in each circuit model via self-organization (through intrinsic plasticity, for example).
- In state of the art NPs, the parameters are stored in volatile memories which lose their information when they are power cycled and hence must be reprogrammed. Therefore, they are obliged to constantly dissipate static power to maintain their information. Thanks to the nonvolatility of resistive memory in hybrid systems, they can be powered on and off without requiring reprogramming and do not require static power to be consumed to retain information.
- Since parameters are remotely set by bias generators, it is required to reprogram the bias generator whenever they are updated. To implement local plasticity mechanisms per model neuron (foregoing the massive power and area drawbacks this would entail), it would be required to read out the firing rate of every neuron and reprogram bias generators per neuron using a "computer in the loop" approach. This imposes a bandwidth limitation resulting from the von Neumann bottleneck that still exists between the distributed model circuits and the centralized digital memory despite the distributed nature of the circuit models themselves. In hybrid systems, the RRAM incorporated into the circuits can be configured locally by additional analog circuits, therefore imposing no limitations on the bandwidth of local plasticity mechanisms.
- A final advantage of using RRAM to determine model parameters, over subthreshold CMOS transistors, is their increased stability under temperature fluctuations. The drain source resistance of transistors biased in the subthreshold regime, as required in neuromorphic processors to realize biological time constants (in the millisecond regime), is famously sensitive to small fluctuations in temperature. This is detrimental for neuromorphic processors since during an application, if the ambient temperature drifts, so will the behavior of the models from those desired. This change in resistance is an exponential function of the ratio of the material activation energy over the temperature change. The measured activation energy of RRAM[28] is an order of magnitude lower than that measured for CMOS transistors[29] and so allows for reduced sensitivity of model circuits on temperature changes. However, for large temperatures, care should be taken over the choice of electrodes. While the electrodes (Ti/TiN) of the technology considered in this paper have demonstrated good retention during temperature cycling (between room temperature up to 200 °C), other

technologies (Pt/Pt electrodes for example) can lose their state completely for increases in temperature.[30]

To quantify the benefits of our approach with respect to the state of the art, we estimate and compare the power and area consumption required by an example state of the art NP vs a system embracing the hybrid approach. On the Dynapse chip,[7] as an example, each bias parameter on average consumes about 4 μW of power. If the chip were to have unique parameters for each neuron, assuming only 3 parameters required per neuron (time constant, gain and refractory period), each neuron would burn 12 μW of power. This power consumption means, with only 1000 individually parameterized neurons, we already burn a hugely undesirable 12 mW of static power. Moreover, to route the 3 aforementioned biases to each neuron from a bias generator, assuming the 4th metal layer in 180 nm technology (same as Dynapse), 1.5 μm^2 of area is required. For 1000 neurons, this number grows to 1.5 mm^2 which is equivalent to half of a whole silicon chip. In comparison, the hybrid approach consumes no static power and needs no routing to bias model circuits.

VI. CONCLUSIONS

In this paper, we proposed that hybrid neuromorphic circuits, those incorporating resistive memories into CMOS neuron and synapse models, can solve a number of problems faced by a fully CMOS approach to neuromorphic processors. Hybrid systems will allow parameter variety and static power consumption to be increased and decreased, respectively, by orders of magnitude and, when compared to deep sub-threshold CMOS neuron and synapse models, the model parameters will exhibit greater stability over an extensive temperature range. Furthermore, the state of the memories can be modified by local circuits in order to implement massively parallel local plasticity mechanisms—currently impossible with existing approaches. In this paper, we explored nonconventional properties of HfO_2 based OxRAM, namely, the stochastic SET operation and the RESET random variable. Using these operations, we proposed and demonstrated a technologically plausible intrinsic plasticity algorithm which allowed DPI neurons interconnected by DPI synapses to realize a recurrent neural network, to self-organize and fire around a target firing rate. The hybrid RNN was able to find a configuration which exhibited the healthy and stable network dynamics required to find use in ultralow power edge-computing problems confronted with data of a temporal nature. Encouragingly, the measured cycle-to-cycle HRS variability was seen to be beneficial for computation, while the intrinsic plasticity algorithm was able to mitigate negative effects of high device-to-device SET probability variability at the expense of longer time to convergence. Like in biology, where there exists a fantastic variety of cell types, resistive memories also come in many flavors and exhibit diverse properties. In addition to the stochastic properties of OxRAM (studied in this paper), the volatile resistive states in silver based conductive bridge RAM can be used to store volatile short term information,[15] while the gradual resistance changes in phase change memories can be used to realize incremental changes in nonvolatile parameters.[31] This work opens up the door to not only the potential of using resistive memories as fundamental building blocks of neuron and synapse models in useful neuromorphic processors but also illustrates why they are a necessity in facilitating future neuromorphic processors to address ultralow power embedded temporal edge-computing problems.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the J. Casas Carnot chair in bio-inspired technologies in addition to that of the NeuRAM3 project.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems* (ACM, 2012), Vol. 1, pp. 1097–1105.
[2] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks **61**, 85–117 (2015).
[3] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in Annual Meeting of the Assocication for Computational Linguistics, 2019.
[4] W. Maass, T. Natschläger, H. Markram, W. Maass, T. Natschlaeger, H. Markram, and W. Maass, "Real-time computing without stable states: A new framework for neural computation based on perturbations," Neural Comput. **14**, 2531–2560 (2002).
[5] T. Dalgaty, E. Vianello, D. Ly, G. Indiveri, B. Salvo, E. Nowak, and J. Casas, "Insect-inspired elementary motion detection embracing resistive memory and spiking neural networks," in *Biomimetic and Biohybrid Systems* (Springer Nature, 2018), pp. 115–128.
[6] T. Dalgaty, E. Vianello, B. De Salvo, and J. Casas, "Insect-inspired neuromorphic computing," Curr. Opin. Insect Sci. **30**, 59–66 (2018).
[7] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)," IEEE Trans. Biomed. Circuits Syst. **12**, 106–122 (2017).
[8] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," Proc. IEEE **102**, 1367–1388 (2014).
[9] A. Lazar, P. Gordon, and J. Triesch, "SORN: A self-organizing recurrent neural network," Front. Comput. Neurosci. **3**, 23 (2009).
[10] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," Nat. Rev. Neurosci. **5**, 97 (2004).
[11] R. Baddeley, L. F. Abbott, M. C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls, "Responses of neurons in primary and inferior temporal visual cortices to natural scenes," Proc. R. Soc. B **264**, 1775–1783 (1997).
[12] M. Stemmler and C. Koch, "How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate," Nat. Neurosci. **2**, 521–527 (1999).
[13] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanović, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, and L. Perniola, "Resistive memories for ultra-low-power embedded computing design," in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), pp. 6.3.1–6.3.4.
[14] A. Grossi, E. Nowak, C. Zambelli, C. Pellissier, S. Bernasconi, G. Cibrario, K. E. Hajjam, R. Crochemore, J. F. Nodin, P. Olivo, and L. Perniola, "Fundamental variability limits of filament-based RRAM," in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 4.7.1–4.7.4.
[15] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, M. Gimzewski, and J. K. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," Nat. Mater. **10**, 591 (2011).
[16] A. Grossi, E. Vianello, C. Zambelli, P. Royer, J. Noel, B. Giraud, L. Perniola, P. Olivo, and E. Nowak, "Experimental investigation of 4-kb RRAM arrays programming conditions suitable for TCAM," IEEE Trans. Very Large Scale Integr. Syst. **26**, 2599–2607 (2018).

[17]T. Dalgaty, "Hybrid CMOS-RRAM neurons with intrinsic plasticity," in *IEEE ISCAS 2019 Conference Proceedings* (IEEE, 2019), Vol. 61.
[18]G. Indiveri and Y. Sandamirskaya, "The importance of space and time in neuromorphic cognitive agents," IEEE Signal Process. Mag. (in press); e-print arXiv:1902.09791.
[19]G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S. C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," Front. Neurosci. **5**, 73 (2011).
[20]D. Heeger, *Markov Chains* (Cambrdige University Press, 2000).
[21]Y. Nishi, U. Bottger, R. Waser, and S. Menzel, "Crossover from deterministic to stochastic nature of resistive-switching statistics in a tantalum oxide thin film," IEEE Trans. Electron Devices **65**, 4320–4325 (2018).
[22]A. Rukhin, J. Soto, J. C. Nechvatal, M. Smid, E. Barker, L. Stefan, and S. Vo, "A statistical test suite for random and pseudorandom number generators for cryptographic applications," NIST: National Institute of Standards and Technology (2010).
[23]D. R. B. Ly, A. Grossi, C. Fenouillet-Beranger, E. Nowak, D. Querlioz, and E. Vianello, "Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning," J. Phys. D: Appl. Phys. **51**, 444002 (2018).
[24]D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo, and B. DeSalvo, "Variability-tolerant convolutional neural network for pattern recognition applications based on oxram synapses," in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), pp. 28.4.1–28.4.4.
[25]G. Piccolboni, G. Molas, D. Garbin, E. Vianello, O. Cueto, C. Cagli, B. Traore, B. De Salvo, G. Ghibaudo, and L. Perniola, "Investigation of cycle-to-cycle variability in HfO_2-based oxram," IEEE Electron Device Lett. **37**, 721–723 (2016).
[26]M. Payvand, M. Nair, L. Muller, and G. Indiveri, "A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: From mitigation to exploitation," Faraday Discuss. **213**, 487 (2018).
[27]J. R. Norris, Poisson Model of Spike Generation (1998).
[28]K. Jung, H. Seo, Y. Kim, H. Im, J. Hong, J.-W. Park, and J.-K. Lee, "Temperature dependence of high- and low-resistance bistable states in polycrystalline NiO films," Appl. Phys. Lett. **90**, 052104 (2007).
[29]V. Obreja and A. Obreja, "Activation energy values from the temperature dependence of silicon pn junction reverse current and its origin," Phys. Status Solidi A **207**, 1252–1256 (2010).
[30]B. Traore, K.-H. Xue, E. Vianello, G. Molas, P. Blaise, B. Salvo, A. Padovani, O. Pirrotta, L. Larcher, L. Fonseca, and Y. Nishi, "Investigation of the role of electrodes on the retention performance of HfOx based RRAM cells by experiments, atomistic simulations and device physical modeling," in *2013 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2013).
[31]O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, "Visual pattern extraction using energy-efficient '2-PCM synapse' neuromorphic architecture," IEEE Trans. Electron Devices **59**, 2206–2214 (2012).